DOCUMENT RESUME

ED 295 962 TM 011 740

AUTHOR Blumberg, Carol Joyce
TITLE Regression Slope Estimation When Both Measurement and Specification Error Are Present.
PUB DATE Apr 88
NOTE 36p.; Paper presented at the Annual Meeting of the American Educational Research Association (New Orleans, LA, April 5-9, 1988).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Computer Simulation; Equations (Mathematics); *Error of Measurement; Graphs; *Least Squares Statistics; *Maximum Likelihood Statistics; *Regression (Statistics); Scores; *Statistical Bias
IDENTIFIERS *Error Analysis (Statistics); Slope Estimation; *Specification Bias; Specification Errors

ABSTRACT

Traditionally, the errors-in-variables problem is concerned with the point estimation of the slope of the true scores regression line when the regressor is measured with error, and when no specification error is present. In this paper, the errors-in-variables problem is extended to include specification error. Least squares procedures provide a biased estimator of the slope of the true scores regression line. Further, the maximum likelihood estimates of the slope (which are consistent) exist only once some assumptions are made. Maximum likelihood estimates are given for the extended version of the errors-in-variables problem (i.e., when specification error is present) under the usual assumptions and under several new assumptions that are more appropriate for the social and behavioral sciences than the previously used assumptions. A simulation study illustrates this process. The results of the study indicate that the maximum likelihood estimates (both under the old and new assumptions) far outperform the least squares procedures when several different criteria (such as bias and standard error) are used. Eleven tables are presented. (Author/TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED295962

REGRESSION SLOPE ESTIMATION WHEN BOTH MEASUREMENT AND SPECIFICATION ERROR ARE PRESENT

Carol Joyce Blumberg

Department of Mathematics and Statistics

Winona State University

Winona, MN 55987

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

X This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

CAROL JOYCE BLUMBERG

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Paper presented at the American Educational Research Association Annual Meeting, New Orleans, April, 1988.

The author is grateful to Andrew C. Porter for his insightful and extremely helpful comments of an earlier version of this paper. The theoretical portions of this paper are heavily based on a paper presented at 1983 Joint Statistical Meetings (Blumberg, 1983).

CG 019 748

ABSTRACT

Traditionally, the errors-in-variables problem is concerned with the point estimation of the slope of the true scores regression line when the regressor is measured with error, and when no specification error is present. In this paper the errors-in-variables problem is extended to also include specification error. As is well known, least squares procedures provide a biased estimator of the slope of the true scores regression line. Further, it is well known the maximum likelihood estimates of the slope (which are consistent) exist only once some assumptions are made. In this paper maximum likelihood estimates are given for the extended version of the errors-in-variables problem (i.e., when specification error is present) under the usual assumptions and under several new assumptions which are more appropriate for the social and behavioral sciences than the previously used assumptions. A simulation study is then described. The results of the simulation study show that the maximum likelihood estimates (both under the old and new assumptions) far outperform the least squares procedures when several different criteria (such as bias and standard error) are used.

REGRESSION SLOPE ESTIMATION WHEN BOTH MEASUREMENT AND SPECIFICATION ERROR ARE PRESENT

One of the problems which has been of great interest over the years to econometricians and others is known as the errors-in-variables problem. Some of the more complete discussions of early work on this problem are M. Brown (1982), Johnston (1972), Kendell & Stuart (1973), Madansky (1959) and Moran (1971). Recently Fuller (1987) has published a volume devoted solely to the errors-in-variable problem that includes the newer, as well as the older, work on this problem. The main purposes of this paper are to present maximum likelihood solutions to the errors-in-variables problem for some situations that have not been investigated by others and to report on a simulation study in which these new methods were studied. For completeness, the maximum likelihood solutions presented in the past literature will also be given.

Definition of the Problem

For ease of discussion, some notation will be introduced presently. Let X_i and Y_i represent two observed variables of interest for individual i . Let X_i^* and Y_i^* represent the values of X_i and Y_i if they were measured without error (i.e., the true scores or latent scores). Let e_{X_i} and e_{Y_i} represent the errors of measurement (i.e., the difference between the true and observed values) in X and Y, respectively. Let α^* and β^* represent the Y^*-intercept and slope, respectively, of the Y^* on X^* regression line. Basically the errors-in-variables problem is concerned with the point estimation of the slope (and Y^*-intercept)

of the Y^* on X^* regression line. Most often discussions of the errors-in-variables problem assume that there exists a correlation of +1 or -1 between X^* and Y^* (M. Brown, 1982; DeGracie & Fuller, 1972; Fuller & Hidiroglou, 1978; Johnson, 1972; Kendall & Stuart, 1973; Madansky, 1959; Moran, 1971; Sprent, 1966; Wald, 1940; Lindley, 1953 and others). That is, most discussions assume that there is no specification error present in the statement of the relationship between X^* and Y^*. The assumption of no specification error is, however, very unrealistic for most, if not all educational, psychological, and other behavioral science applications. Further, the assumption of no specification error is necessry. Hence, the definition of the errors-in-variables problem will be extended in this paper to include specification error. A fairly thourough search of the literature revealed few sources where specification errors were considered. The only sources found were Cochran (1968) and Rock, Werts, Linn, and Jöreskog (1977). Neither Cochran nor Rock, et. al., however, discuss explicit solutions to the errors-in-variables problem when specification error is added. Cochran only talks about the errors-in-variables problem in relationship to ANCOVA and Rock et. al. only talk in general terms about how, once enough identifying or overidentifying restrictions are made, LISREL (Jöreskog & Sörbom, 1981) can be used to obtain the values of the desired maximum likelihood estimates.

The errors-in-variables problem can then be expressed symbolically as follows:

Find a consistent point estimator of β^* when

$$Y_i^* = \alpha^* + \beta^* \cdot X_i^* + e_{s_i} \quad , \tag{1}$$

$$X_i = X_i^* + e_{X_i} \quad , \tag{2}$$

and

$$Y_i = Y_i^* + e_{Y_i} \quad , \tag{3}$$

where e_{s_i} is the specification error for individual i.

Further, it is assumed that the vector $(X^*, Y^*, e_X, e_Y, e_s)$ has a multivariate normal distribution with mean vector $(\mu_X, \mu_Y, 0, 0, 0)$ and with a variance-covariance matrix all of whose off-diagonal elements (i.e., the covariances) except $\sigma_{X^*Y^*}$ (the covariance of X^* and Y^*) are zero and whose diagonal elements (i.e., the variances) are in order symbolized by $\sigma^2_{X^*}$, $\sigma^2_{Y^*}$, $\sigma^2_{e_X}$, $\sigma^2_{e_Y}$, and $\sigma^2_{e_s}$.

Background

There are many educational and other behavioral science situations in which the estimation of β^* is desirable. The author perused several recent years of American Educational Research Journal and of Journal of Educational Research. Many examples were encountered where the authors estimated β^* for a variety of variables, X and Y. All, however, used the least squares estimator of β^*.

The usual least squares estimate of β, the slope of the Y on X regression line when X is measured without error, is given by $\hat{\beta}_{LS} = \frac{S_{XY}}{S_X^2}$, where S_{XY} and S_X^2 are the sample covariance of

X and Y and the variance of Y, respectively. It is well known that $\hat{\beta}_{LS}$ is not a consistent estimator of β^*, since $E(\hat{\beta}_{LS}) = \rho_{XX} \cdot \beta^*$ (Berkson, 1950; Johnston, 1972; Lindley, 1947; and others). Hence, it has been suggested that maximum likelihood estimation or other techniques be used instead. In order to compute the maximum likelihood estimates of α^* and β^*, it is necessary to get expressions for the population means, variances, and covariances of the observed variables in terms of β^* and other parameters of interest. Hence

$$\mu_X = \mu_X \quad ; \tag{4}$$

$$\mu_Y = \alpha^* + \beta^* \cdot \mu_X \quad , \text{ from equation (1)} \; ; \tag{5}$$

$$\sigma_X^2 = \sigma_X^{2*} + \sigma_{e_X}^2 \quad , \text{ from equation (2)} \; ; \tag{6}$$

$$\sigma_Y^2 = (\beta^*)^2 \cdot \sigma_X^{2*} + \sigma_{e_Y}^2 + \sigma_{e_s}^2 \quad , \tag{7}$$

from equations (1) and (3);

and

$$\sigma_{XY} = \beta^* \cdot \sigma_X^{2*} \quad , \tag{8}$$

where σ_{XY} is the population covariance of X and Y.

Equation (8) can be derived as follows:

$$\sigma_{XY} = \text{Cov}(X,Y)$$

$$= \text{Cov}(X^* + e_X, Y^* + e_Y) \qquad \text{[from equations (2) and (3)]}$$

$$= \text{Cov}(X^*,Y^*) + \text{Cov}(X^*,e_Y) + \text{Cov}(e_X,Y^*) + \text{Cov}(e_X,e_Y) \quad .$$

Next, by assumption, $Cov(X^*,e_Y)$, $Cov(e_X,Y^*)$, and $Cov(e_X,e_Y)$ are all zero. Hence $\sigma_{XY} = Cov(X^*,Y^*)$. Therefore,

$$\begin{aligned}\sigma_{XY} &= Cov(X^*,Y^*) \\ &= Cov(X^*,\alpha^* + \beta^*.X^* + e_s) \quad \text{from equation (1)} \\ &= Cov(X^*,\alpha^*) + Cov(X^*,\beta^*\cdot X^*) + Cov(X^*,e_s) \\ &= \beta^*\cdot Cov(X^*,X^*) \ ,\end{aligned}$$

since $Cov(X^*,e_s)$ is zero by assumption and since $Cov(X^*,\alpha^*) = 0$, because α^* is a constant. Consequently, $\sigma_{XY} = \beta^*\cdot\sigma_X^{2*}$. Equations (6) and (7) are derived similarly.

As is widely known (e.g. Kendall & Stuart, 1973; Mood, Graybill, & Boes, 1974), the maximum likelihood estimates of μ_X, μ_Y, σ_X^2, σ_Y^2, and σ_{XY} are given by $\bar{X}$, $\bar{Y}$, S_X^2, S_Y^2, and S_{XY}, respectively, when it is assumed that the joint distribution of X and Y is multivariate normal. Theoretically, the maximum likelihood estimates of the unknown parameters (i.e., α^*, β^*, μ_X, σ_X^{2*}, $\sigma_{e_X}^2$, and $\sigma_{e_Y}^2$) on the right hand side of the system of equations (4) to (8) are then computed by solving the system for these unknown parameters in terms of the maximum likelihood estimates of μ_X, μ_Y, σ_X^2, σ_Y^2, and σ_{XY} (Mood, Graybill, & Boes, 1974). But, since there are only five equations in six unknowns, an infinite number of solutions exist. The presence of an infinite number of solutions is not, however, of

practical use. Hence, it is necessary to make additional assumptions so that an infinity of solutions do not exist.

Previous Solutions

Some assumptions that have been made in the past literature are:

Assumption A: $\sigma^2_{e_X}$ is known and $\sigma^2_{e_s} = 0$;

Assumption B: $\sigma^2_{e_Y}$ is known and $\sigma^2_{e_s} = 0$;

Assumption C: The ratio $\frac{\sigma^{2*}_X}{\sigma^2_{e_X}}$ is known and $\sigma^2_{e_s} = 0$;

and

Assumption D: The ratio $\frac{\sigma^2_{e_Y}}{\sigma^2_{e_X}}$ is known and $\sigma^2_{e_s} = 0$.

Assumption C is equivalent to assuming that the reliability of X is known. The reliability of a measure, say X, is defined as the ratio of true score variance to observed score variance and is denoted by ρ_{XX}. To see the equivalence of Assumption C and ρ_{XX} being known, let $k = \frac{\sigma^{2*}_X}{\sigma^2_{e_X}}$. Then $\rho_{XX} = \frac{\sigma^{2*}_X}{\sigma^2_X} =$

$\frac{\sigma^{2*}_X}{\sigma^{2*}_X + \sigma^2_{e_X}} = \frac{k}{k+1}$, by dividing the numerator and denominator by $\sigma^2_{e_X}$. The maximum likelihood estimates of β^* under

Assumptions A, B, C, and D are:

Under Assumption A, $\hat{\beta}^* = \dfrac{S_{XY}}{S_X^2 - \sigma_{e_X}^2}$ (Kendall & Stuart, 1973). (9)

Under Assumption B, $\hat{\beta}^* = \dfrac{S_Y^2 - \sigma_{e_Y}^2}{S_{XY}}$ (Kendall & Stuart, 1973).

Under Assumption C, $\hat{\beta}^* = \dfrac{(k+1)\cdot S_{XY}}{k\cdot S_X} = \dfrac{S_{XY}}{\rho_{XX}\cdot S_X^2} = \dfrac{1}{\rho_{XX}}\cdot\hat{\beta}_{LS}$ (10)

(Johnston, 1972).

Under Assumption D, $\hat{\beta}^* = \dfrac{S_Y^2 - \lambda\cdot S_X^2 + \sqrt{(S_Y^2 - \lambda\cdot S_X^2)^2 + 4\lambda\cdot(S_{XY})^2}}{2\cdot S_{XY}}$

(Kendall & Stuart, 1972). Under all of these assumptions and under the new assumptions to be discussed later, $\hat{\alpha} = \bar{Y} - \hat{\beta}^*\cdot\bar{X}$. For Assumption A, when an estimate of $\sigma_{e_X}^2$ is available instead of $\sigma_{e_X}^2$ being known, DeGracie and Fuller (1972) have derived a consistent non-maximum likelihood estimator of β^*.

Interestingly, the assumption that $\sigma_{e_s}^2 = 0$ is not needed in order to obtain the maximum likelihood estimates of β^* when either $\sigma_{e_X}^2$ is known (i.e., Assumption A) or when ρ_{XX} is known (i.e., Assumption C). The reason is that if one sets the value

of $\sigma^2_{e_s}$ equal to some arbitrary constant the maximum likelihood estimates of β^* can still be calculated in both these cases, since setting $\sigma^2_{e_s}$ equal to some arbitrary value allows the system of equations (4) to (8) to be solved for β^*. Further, no matter what this arbitrary value of $\sigma^2_{e_s}$ is, the maximum likelihood estimate of β^* is the same as in equation (9) [when $\sigma^2_{e_X}$ is known] or as in equation (10) [when ρ_{XX} is known]. Since these estimates do not change as $\sigma^2_{e_s}$ changes, the maximum likelihood estimates in these two cases when $\sigma^2_{e_s}$ is unknown are still given by equation (9) and equation (10).

New Solutions

Before giving the maximum likelihood solutions to the errors-in-variables problem under situations which have not been discussed in the past literature, a listing of the possible assumptions that can made concerning the parameters $\sigma^2_{e_X}$, $\sigma^2_{e_Y}$, $\sigma^2_{e_s}$, and σ^{2*}_X will be given. These possible assumptions are:

Assumption 1: $\sigma^2_{e_X}$ is known ,

Assumption 2: $\sigma^2_{e_Y}$ is known ,

Assumption 3: The ratio $\frac{\sigma^{2*}_X}{\sigma^2_{e_X}}$ is known (i.e., ρ_{XX} is known),

Assumption 4: The ratio $\frac{\sigma^2_{e_Y}}{\sigma^2_{e_X}}$ is known ,

Assumption 5: $\sigma^2_{e_s}$ is known ,

Assumption 6: $\sigma^2_{e_s} + \sigma^2_{e_Y}$ is known ,

Assumption 7: The ratio $\frac{\sigma^2_{e_s} + \sigma^2_{e_Y}}{\sigma^2_{e_X}}$ is known ,

Assumption 8: The ratio $\frac{\sigma^2_{e_s}}{\sigma^2_{e_Y}}$ is known,

Assumption 9: The ratio $\frac{\sigma^2_{e_s}}{\sigma^2_{e_X}}$ is known,

Assumption 10: $\sigma^2_X{}^*$ is known,

Assumption 12: ρ_{YY} is known (i.e., the ratio $\frac{\sigma^2_Y}{\sigma^2_{e_Y}}$ is known),

and

Assumption 11: $\rho_{XX} = \rho_{YY}$.

It is realized that some of these assumptions are more plausible than others. The relative plausibility of the various assumptions will differ across data analysis situations. It is interesting to note that all of the assumptions except Assumption 11 actually require knowledge on the part of the data analyst. Assumption 11, however, only requires believa-

ven though Assumptions 1 to 10, & 12 are stated in

———— is known, it is realized that in actual

data analysis situations the analysts will be providing their best guess at the correct value of a particular quantity rather than knowing the correct value. In the remainder of this paper it will be assumed nevertheless that the correct values of particular quantities are known rather than guessed at.

The solutions under Assumption 1 or under Assumption 3 have already been discussed when it was shown that the $\sigma^2_{e_s} = 0$ part of Assumptions A and C is not necessary. Solutions can also be derived when either Assumption 6, 7, or 10 is used. When any of the remaining assumptions (i.e., Assumptions 2, 4, 5, 8, 9, 11, or 12) is taken singly, an infinite number of solutions for β^*, in terms of σ^2_X, σ^2_Y, σ_{XY}, and the particular assumption taken, still exist. Hence, maximum likelihood estimates of β^* are not available when any of the Assumptions 2, 4, 5, 8, 9, 11, or 12 is taken singly. Solutions can be derived, however, when any combination of two of these assumptions, except 4 and 11 or 2 and 12, is jointly assumed. When Assumptions 4 and 11 or Assumptions 2 and 12 are jointly assumed an infinite number of solutions for β^* still exist, and hence maximum likelihood estimates still do not exist. Only elementary algebra is necessary to derive the maximum likelihood estimates, even though in some cases the derivations are quite tedious. Hence the author has chosen in this paper to give only the maximum likelihood estimates and eliminate the derivations of these estimates.

Table 1 gives these maximum likelihood estimates of β^* under the various assumptions. In this table when more than one set of assumptions are listed together in the first column (e.g., 3 alone; 11 and 12 jointly), this means that each of these sets of assumptions generates the same solution for $\hat{\beta}^*$. When two or more expressions for β^* are given together in the second column and are connected by a ; , this means that these expressions are two different "computational" formulae for the same estimator.

Insert Table 1 about here

Set-Up of Simulation Study

A simulation study was performed that involved 1000 data sets at each of three levels of data reliability crossed with three levels of sample sizes. This simulation study is only a first step in the comparison of the various estimators proposed in the last section. As will be discussed later, more remains to be done. In order to be able to compare the properties of the various estimators, the generated data sets simultaneously met the Assumptions 1 through 12. For all data sets, without loss of generality, $\sigma^2_{X^*}$ was set to 1, and then $\sigma^2_{e_s}$ was set to 1/4. Since, as was stated earlier, no previous work is known to the author where specification error was included, the choice of $\sigma^2_{e_s}$ as being equal to 1/4 of $\sigma^2_{X^*}$ was made because even in those social and behavioral science situations where errors of measurement are negligible, and where a linear relationship makes theoretical sense, the r_{XY} 's seem to often be around .75,

leaving approximately one-quarter of the variance being due to specification error. For this simulation study the values of $\alpha^* = 0$. Making $\alpha^* = 0$ was done in order to be able to focus on the estimation of β^*. Arbitrarily, β^* was set equal to 1.5 . This simulation study needs, however, be repeated in the future with different values of β^* . Once β^* is set equal to 1.5 and $\sigma_X^{2*} = 1$, it follows that $\sigma_Y^{2*} = 2.25$.

The three levels of data reliability chosen for this study were .5, .7, and .9 . In order to be able to compare across the various assumptions, each reliability value represents the common value of ρ_{XX} and ρ_{YY} . Once the reliability value is chosen the remainder of simulation population values can be explicitly determined since they are functions of β^* , σ_X^{2*} , ρ_{XX} , and ρ_{YY} . These values are:

Parameter	Value when $\rho = .5$	Value when $\rho = .7$	Value when $\rho = .9$
$\sigma^2_{e_X}$	1	3/7	1/9
$\sigma^2_{e_Y}$	9/4	27/28	1/4

The sample sizes chosen for this study were 25, 50, and 100, since these are typical sample sizes in the social and behavioral sciences.

The first stage of the simulation study was to generate pseudo-random standard normal deviates. The second stage of the simulation involved estimating the Estimators 1 through 15 and the usual least squares estimator for 1000 data sets based on pairs of observations, X and Y, that were generated using the pseudo-random deviates generated at the first stage. For each

pair of observations, X and Y, (i.e., simulated individuals) four standard normal deviates were needed. The first deviate represented X^*. The second and third deviates were multiplied by the appropriate values to give e_X and e_Y. The fourth deviate was multiplied by .25 to give e_S. Each deviate was generated separately by entering four different seeds simultaneously into the RANNOR pseudo-random standard normal generator in SAS Version 5.1 as implemented on the VAX computer at Winona State University running under the VMS operating system. Due to internal limitations on VAX at Winona State, only 6250 individuals' deviates could be generated per SAS run. Four new seeds were entered for each of the 16 runs used to generate the 400,000 deviates needed in this study (1000 simulation runs x 100 individuals x 4 deviates per individual). The RANNOR generator in SAS worked well. The means for the 64 sets of 6250 pseudo-standard normal deviates ranged from a low of -.03833002 to a high of +.01904790 while the standard deviations ranged from a low of .98150919 to a high of 1.01668405 . The pairwise correlations between the four deviates within a run ranged from a low of -.02693 to a high of +.03227 .

Results

For the reliability values of .5 and .7 all of the maximum likelihood estimators studied did better than the usual least squares estimator, while for the reliability value of .9 the maximum likelihood estimators did better in almost all cases. Tables 3 through 11 provide summaries of the simulation results

for the various reliability values and sample sizes. Table 2 provides the key to the estimator numbers used in Tables 3 through 11. In each table the last column represents the

---

Insert Tables 2 through 11 about here

---

percentage of times that the estimator being studied is closer to the correct theoretical value of 1.5 for β^* than the usual least squares estimator is. Although this is not a traditional method for comparing estimators, it was felt by the author that this column gives extremely valuable information in this situation. The column labeled "Mean" in each of Tables 3 to 11 provides the observed means for each estimator rather than the theoretical values, even though the theoretical values can easily be derived. It should be noted that the theoretical values were calculated by the author and the observed values for the mean were indeed close to the theoretical values. The reason, however, that a simulation study was done was to obtain estimates of the standard errors of the various estimators and to assess the percentage of times each estimator outperformed the usual least squares estimator. Since observed standard errors and percentages are reported in the tables, the observed means are reported simply for consistency purposes.

The first thing that should be noticed from Tables 3 to 11 is, as would be expected, that as the sample sizes increase within a reliability value all of the estimators, with only a few exceptions, have decreasing standard errors, less bias (i.e., the observed means become closer to the theoretical value of 1.5),

and hence increased percentages of times better than least squares estimator. Estimators 10, 12, and 13 for the reliability of .5; Estimator 5 for the reliability value of .7; and Estimator 5 for the reliability value of .9 are the exceptions and they are exceptions only for the bias of the mean. It should be noticed, however, that these estimators have very little bias already for samples of size 25. Hence, this nondecrease in the values for the means could be caused simply by nested sets of random numbers being used for the simulations at the different sample sizes (i.e., the 100,000 random deviates used for n = 25 are a subset of the 200,000 random deviates used for n = 50, which in turn are a subset of the 400,000 deviates used for n = 100.)

The second thing to notice is that within a sample size (i.e., n = 25, n = 50, or n = 100) in all cases the percentage of times that an estimator does better than the usual least squares estimator decreases as the reliability of the data increases. This is not all that surprising when one remembers that the bias in the least squares estimator of β^* is heavily dependent on the reliability of the data. That is, the bias of the least squares estimator is equal to $(\rho_{XX} - 1)\cdot\beta^*$ and hence the bias increases as the reliability of the data decreases. Further, as would be expected, within a sample size almost always the bias and the standard errors of the various estimators decrease as the reliability increases. The exceptions are the bias for Estimators 10, 12, and 13 for the sample sizes of n = 25 and 50 and Estimators 2, 3, 10, 12, and 13 when n = 100.

Conclusions

A final thing to notice across the estimators is that all of the estimators except Estimators 3, 5, and 9 have similar standard deviations and percentage of times better than the usual least squares estimator for a fixed sample size and reliability level (except in the case when n = 25 and the reliability value is .5). Hence, when several of the assumptions made in generating the various estimators seem reasonable simultaneously in a particular real-world setting, this simulation study does not give very much advice for choosing between the various estimators. What can be concluded, however, is that if a variety of assumptions fit the real-world setting, then the researcher should avoid using Estimators 3, 5, and 9. On the other hand, and more importantly, all of the estimators presented here, except Estimator 9, do much better than the usual least squares estimator. Hence, as long as one of the sets of assumptions (except those that lead to Estimator 9) presented in this paper holds, the corresponding maximum likelihood estimators should be used over the usual least squares estimator. As mentioned above, the only estimator that performed badly was Estimator 9. The author can not account for this. Perhaps there was an algebraic error or a programming error on the author's part, but after several checks by the author neither type of error was detected.

Limitations and Future Directions

There are two major limitations to this study. First, only one value of β^* was studied and only one value of $\sigma^2_{e_s}$ was used when studying this value of the true scores regression line slope. Hence, further simulation studies need to be performed where the values of β^* and $\sigma^2_{e_s}$ are systematically varied. Second, this study was performed using only data sets where all of the assumptions were met simultaneously. A study needs to be done where different assumptions are violated and the effects of these violations are studied for the various estimators.

References

Berkson, J. Are there two regressions? Journal of the American Statistical Association, 1950, 45, 164-180.

Blumberg, C. J. Errors-in-variables: Maximum Likelihood solutions under reasonable behavoral science assumptions. American Statistical Association 1983 Proceedings of the Social Statistics Section, 1983, 109-114.

Brown, G. H. Generalized least squares applied to the linear ultrastructural model. Biometrika, 1978, 65, 441-444.

Brown, M. L. Robust line estimation with errors in both variables. Journal of the American Statistical Association, 1982, 77, 71-79

Cochran, W. G. Errors of measurement in statistics. Technometrics, 1968, 10, 637-666.

DeGracie, J. S., & Fuller, W. A. Estimation of the slope and analysis of covariance when the concomitant variable is measured with error. Journal of the American Statistical Association, 1972, 67, 930-937.

Fuller, W. A. Measurement error models. New York: John Wiley & Sons, 1987.

Fuller, W. A., & Hidiroglou, M. A. Regression estimation after correcting for attenuation. Journal of the American Statistical Association, 1978, 73, 99-104.

Johnston, J. Econometric methods, (2nd edition). New York: McGraw-Hill, 1972.

Kendall, M. G., & Stuart, A. The advanced theory of statistics, Volume 2, Inference and relationship (3rd edition). London: Griffin, 1973.

Jöreskog, K. G., & Sörbom, D. LISREL V: User's guide. Chicago: International Educational Services, 1981.

Lindley, D. V. Regression lines and the linear functional relationship. Supplement Journal of the Royal Statistical Society, 1947, 9, 218-244.

Lindley, D. V. Estimation of a functional relationship. Biometrika, 1953, 40, 47-49.

Madansky, A. The fitting of straight lines when both variables are subject to error. Journal of the American Statistical Association, 1959, 54, 173-205.

Mood, A. M., Graybill, F. A., & Boes, D. C. Introduction to the theory of statistics, (3rd edition). New York: McGraw-Hill, 1974.

Moran, P. A. P. Estimating structural and functional relationships. Journal of Multivariate Analysis, 1971, 1, 231-255.

Patel, J. K., Kapadia, C. H., & Owen, D. B. Handbook of Statistical Distributions. New York: Marcel Dekker, 1976.

Rock, D. A., Werts, C. E., Linn, R. L., & Jöreskog, K. G. A maximum likelihood solution to the errors in variables and errors in equations model. The Journal of Multivariate Behavioral Research, 1977, 12, 187-197.

Spiegelman, C. On estimating the slope of a straight line when both variagles are subject to error. The Annals of Statistics, 1979, 7, 201-206.

Sprent, P. A generalized least-squares approach to linear functional relationships. Journal of the Rpyal Statistical Society, Series B, 1966, 28, 278-297.

Wald, A. The fitting of straight lines if both variables are subject to error. Annals of Mathematical Statistics, 1940, 11, 284-300.

Table 1

Maximum Likelihood Estimates of β^*
Under Various Assumptions

Assumptions	Maximum Likelihood Estimate of β^* ($\hat{\beta}^*$)
1	$\dfrac{S_{XY}}{S_X^2 - \sigma^2_{e_X}}$
3 or 11 & 12 jointly	$\dfrac{S_{XY}}{\rho_{XX} \cdot S_X^2}$; $\dfrac{\varepsilon + 1}{\varepsilon} \cdot \dfrac{S_{XY}}{S_X^2}$ where $\varepsilon = \dfrac{\sigma^{2*}_X}{\sigma^2_{e_X}}$
6 or 2 & 5 jointly or 2 & 8 jointly or 5 & 8 jointly	$\dfrac{S_Y^2 - (\sigma^2_{e_Y} + \sigma^2_{e_S})}{S_{XY}}$
7 or 4 & 8 jointly or 4 & 9 jointly or 8 & 9 jointly	$\dfrac{S_Y^2 - \eta \cdot S_X^2 + \sqrt{(S_Y^2 - \eta \cdot S_X^2)^2 + 4\eta (S_{XY})^2}}{2 \cdot S_{XY}}$ where $\eta = \dfrac{\sigma^2_{e_S} + \sigma^2_{e_Y}}{\sigma^2_{e_X}}$
10	$\dfrac{S_{XY}}{\sigma^{2*}_X}$

continued on next page

Table 1 (continued)

Assumptions	Maximum Likelihood Estimate of β^* ($\hat{\beta}^*$)
2 & 4 jointly	$\frac{S_{XY}}{S_X^2 - \sigma_{e_X}^2}$; $\frac{S_{XY}}{S_X^2 - \frac{\sigma_{e_Y}^2}{\lambda}}$ where $\lambda = \frac{\sigma_{e_Y}^2}{\sigma_{e_X}^2}$
2 & 9 jointly	$\frac{S_Y^2 - \sigma_{e_Y}^2 - \delta S_X^2 + \sqrt{(S_Y^2 - \sigma_{e_Y}^2 - \delta S_X^2)^2 + 4\delta (S_{XY})^2}}{2S_{XY}}$ where $\delta = \frac{\sigma_{e_s}^2}{\sigma_{e_X}^2}$
2 & 11 jointly	$\frac{S_{XY} \cdot S_Y^2}{S_X^2 \cdot (S_Y^2 - \sigma_{e_Y}^2)}$
4 & 5 jointly	$\frac{S_Y^2 - \sigma_{e_s}^2 - \lambda S_X^2 + \sqrt{(S_Y^2 - \sigma_{e_s}^2 - \lambda S_X^2)^2 + 4\lambda (S_{XY})^2}}{2S_{XY}}$ where $\lambda = \frac{\sigma_{e_Y}^2}{\sigma_{e_X}^2}$
4 & 12 jointly	$\frac{S_{XY}}{S_X^2 - \lambda(1 - \rho_{YY})S_Y^2}$ where $\lambda = \frac{\sigma_{e_Y}^2}{\sigma_{e_X}^2}$
5 & 9 jointly	$\frac{S_{XY}}{S_X^2 - (\sigma_{e_X}^2)}$; $\frac{S_{XY}}{S_X^2 - \frac{\sigma_{e_s}^2}{\delta}}$ where $\delta = \frac{\sigma_{e_s}^2}{\sigma_{e_X}^2}$

continued on next page

Table 1 (continued)

Assumptions	Maximum Likelihood Estimate of β^* $(\hat{\beta}^*)$
5 & 11 jointly	$\dfrac{-\sigma^2_{e_s} + \sqrt{(\sigma^2_{e_s})^2 + 4 \cdot \dfrac{S_Y^2 \cdot (S_{XY})^2}{S_X^2}}}{2 \cdot S_{XY}}$
5 & 12 jointly	$\dfrac{\rho_{YY} \cdot S_Y^2 - \sigma^2_{e_s}}{S_{XY}}$
8 & 11 jointly	$\dfrac{-\gamma \cdot S_Y^2 + \sqrt{(\gamma \cdot S_Y^2)^2 + 4 \cdot (1+\gamma) \cdot \dfrac{S_Y^2 \cdot (S_{XY})^2}{S_X^2}}}{2 \cdot S_{XY}}$ where $\gamma = \dfrac{\sigma^2_{e_s}}{\sigma^2_{e_Y}}$
9 & 11 jointly	$\dfrac{-\delta \cdot S_X^2 + \sqrt{(\delta \cdot S_X^2)^2 + 4 \cdot \left(\dfrac{S_Y^2}{S_X^2} + \delta\right) \cdot (S_{XY})^2}}{2 \cdot S_{XY}}$ where $\delta = \dfrac{\sigma^2_{e_s}}{\sigma^2_{e_X}}$

continued on next page

Table 1 (continued)

Assumptions	Maximum Likelihood Estimates of β^* $(\hat{\beta}^*)$
8 & 12 jointly	$\frac{[\rho_{YY} + \gamma \cdot (\rho_{YY} - 1)] \cdot S_Y^2}{S_{XY}}$ where $\gamma = \frac{\sigma^2_{e_s}}{\sigma^2_{e_v}}$
9 & 12 jointly	$\frac{\rho_{YY} \cdot S_Y^2 - \delta \cdot S_X^2 + \sqrt{(\rho_{YY} \cdot S_Y^2 - \delta \cdot S_X^2)^2 + 4\delta (S_{XY})^2}}{2 \cdot S_{XY}}$ where $\delta = \frac{\sigma^2_{e_s}}{\sigma^2_{e_X}}$
2 & 12 jointly or 4 & 11 jointly	No maximum likelihood estimates exist (not enough information)

TABLE 2

Correspondence Between Estimator Numbers and Assumptions

Estimator #	Assumptions
1	1; 2 & 4 jointly; 5 & 9 jointly
2	3; 11 & 12 jointly
3	6; 2 & 5 jointly; 2 & 8 jointly; 5 & 8 jointly
4	7; 4 & 8 jointly; 4 & 9 jointly; 8 & 9 jointly
5	10
6	2 & 9 jointly
7	2 & 11 jointly
8	4 & 5 jointly
9	4 & 12 jointly
10	5 & 11 jointly
11	5 & 12 jointly
12	8 & 11 jointly
13	9 & 11 jointly
14	8 & 12 jointly
15	9 & 12 jointly

TABLE 3

Maximum Likelihood Estimates
Using n = 25 and ρ = .5

Estimator Number	Mean	Standard Error	Percentage of Times Better Than Least Squares
Least Squares	0.78739	0.31837	———
1	2.32623	6.05184	63.4%
2	1.57478	0.63674	75.8%
3	1.73922	3.66971	58.3%
4	1.77828	2.35811	75.9%
5	1.51421	0.72259	75.4%
6	1.75161	3.40772	66.0%
7	2.02469	3.95133	70.3%
8	1.96543	1.84852	68.0%
9	*	*	*
10	1.51160	0.38382	88.5%
11	1.90996	2.69919	69.8%
12	1.51126	0.37071	88.7%
13	1.51479	0.40089	87.9%
14	1.94126	2.81998	69.2%
15	1.86877	2.57002	76.1%

Note: The author made a formatting error when printing out the values of Estimator 9 for the 1000 generated data sets. This error was caused by the estimator taking on values less than -9.999995 or greater than 99.999995 . Hence, summary statistics are not reported here for Estimator 9.

TABLE 4

Maximum Likelihood Estimates
Using n = 50 and ρ = .5

Estimator Number	Mean	Standard Error	Percentage of Times Better Than Least Squares
Least Squares	0.77055	0.21214	———
1	1.81852	1.91392	76.7%
2	1.54109	0.42427	86.8%
3	1.54068	0.75099	79.3%
4	1.59450	0.60344	86.6%
5	1.51537	0.50059	85.2%
6	1.55717	0.70521	84.9%
7	1.64410	0.48278	87.3%
8	1.79726	0.71365	79.4%
9	2.97829	4.88477	52.9%
10	1.51545	0.21600	96.0%
11	1.62225	0.67088	89.5%
12	1.51488	0.20835	96.4%
13	1.51639	0.23032	95.7%
14	1.63041	0.67880	89.0%
15	1.62036	0.62184	92.0%

TABLE 5

Maximum Likelihood Estimates
Using n = 100 and ρ = .5

Estimator Number	Mean	Standard Error	Percentage of Times Better Than Least Squares
Least Squares	0.75573	0.13974	———
1	1.61274	0.49832	88.4%
2	1.51146	0.27947	95.9%
3	1.51250	0.41903	93.2%
4	1.54117	0.31076	95.3%
5	1.49434	0.34833	94.5%
6	1.52162	0.38473	94.8%
7	1.56125	0.25456	96.9%
8	1.73606	0.35537	89.3%
9	2.68236	3.99031	58.0%
10	1.50831	0.14386	99.3%
11	1.55318	0.31720	97.4%
12	1.50822	0.13731	99.4%
13	1.50912	0.15131	99.3%
14	1.55724	0.31735	97.5%
15	1.55315	0.29047	98.3%

TABLE 6

Maximum Likelihood Estimates
Using n = 25 and ρ = .7

Estimator Number	Mean	Standard Error	Percentage of Times Better Than Least Squares
Least Squares	1.09879	0.28160	———
1	1.74903	0.91796	64.9%
2	1.56970	0.40228	69.9%
3	1.51658	0.61446	62.6%
4	1.59520	0.50276	72.1%
5	1.50493	0.55060	56.8%
6	1.54858	0.55868	72.2%
7	1.62745	0.32913	69.4%
8	1.67538	0.62468	69.0%
9	2.30126	3.42575	55.3%
10	1.53710	0.26942	78.7%
11	1.60173	0.82520	68.8%
12	1.54076	0.26216	77.6%
13	1.54411	0.28501	76.8%
14	1.62160	0.89816	67.7%
15	1.61040	0.78194	76.0%

TABLE 7

Maximum Likelihood Estimates
Using n = 50 and ρ = .7

Estimator Number	Mean	Standard Error	Percentage of Times Better Than Least Squares
Least Squares	1.07564	0.18404	———
1	1.59831	0.35701	77.1%
2	1.53663	0.26290	81.8%
3	1.49710	0.32149	79.8%
4	1.54091	0.27622	82.6%
5	1.50480	0.37540	75.0%
6	1.51384	0.29097	84.6%
7	1.56004	0.20159	85.0%
8	1.61502	0.28937	78.4%
9	1.86029	0.69816	61.6%
10	1.51886	0.18349	87.8%
11	1.52707	0.27130	86.3%
12	1.52117	0.17718	87.8%
13	1.52254	0.19391	86.9%
14	1.53406	0.27147	86.6%
15	1.53381	0.24538	88.1%

TABLE 8

Maximum Likelihood Estimates
Using $n = 100$ and $\rho = .7$

Estimator Number	Mean	Standard Error	Percentage of Times Better Than Least Squares
Least Squares	1.05882	0.11967	———
1	1.54251	0.21049	89.4%
2	1.51260	0.17097	93.5%
3	1.49829	0.22088	92.2%
4	1.51980	0.17851	93.0%
5	1.49121	0.26691	89.3%
6	1.50696	0.19677	93.8%
7	1.52788	0.12628	95.8%
8	1.59659	0.18871	88.7%
9	1.74638	0.35558	69.8%
10	1.50877	0.12275	97.0%
11	1.51632	0.17821	95.8%
12	1.50986	0.11724	97.1%
13	1.51072	0.12821	96.9%
14	1.52053	0.17641	96.6%
15	1.51936	0.16030	96.5%

TABLE 9

Maximum Likelihood Estimates
Using n = 25 and ρ = .9

Estimator Number	Mean	Standard Error	Percentage of Times Better Than Least Squares
Least Squares	1.40986	0.20824	———
1	1.59323	0.25242	52.4%
2	1.56651	0.23136	54.6%
3	1.49154	0.22689	52.3%
4	1.55339	0.21167	56.3%
5	1.50040	0.45821	28.7%
6	1.53634	0.20634	59.3%
7	1.57857	0.20509	53.8%
8	1.55575	0.20939	57.1%
9	1.63639	0.29926	48.6%
10	1.53345	0.18688	60.5%
11	1.50881	0.21678	55.6%
12	1.54617	0.18711	58.0%
13	1.54838	0.19755	57.4%
14	1.52436	0.22482	57.1%
15	1.54374	0.20053	60.1%

TABLE 10

Maximum Likelihood Estimates
Using n = 50 and ρ = .9

Estimator Number	Mean	Standard Error	Percentage of Times Better Than Least Squares
Least Squares	1.38100	0.13626	———
1	1.54664	0.16308	62.3%
2	1.53445	0.15140	64.2%
3	1.49482	0.15243	62.3%
4	1.52760	0.14528	64.7%
5	1.49769	0.31291	33.9%
6	1.51866	0.14212	66.4%
7	1.54020	0.13428	64.0%
8	1.53799	0.14339	63.7%
9	1.59003	0.19261	55.5%
10	1.51740	0.12884	67.6%
11	1.50245	0.13856	64.7%
12	1.52409	0.12758	66.3%
13	1.52523	0.13565	65.5%
14	1.50932	0.13476	65.2%
15	1.52176	0.13467	67.0%

TABLE 11

Maximum Likelihood Estimates
Using $n = 100$ and $\rho = .9$

Estimator Number	Mean	Standard Error	Percentage of Times Better Than Least Squares
Least Squares	1.36317	0.08547	———
1	1.52139	0.10001	74.1%
2	1.51463	0.09500	74.8%
3	1.49743	0.10769	75.4%
4	1.51322	0.09513	76.6%
5	1.48945	0.22639	46.4%
6	1.50909	0.09564	76.6%
7	1.51856	0.08238	75.9%
8	1.52881	0.09681	72.8%
9	1.56321	0.12001	63.5%
10	1.50804	0.08598	78.1%
11	1.50233	0.09719	77.1%
12	1.51116	0.08287	78.3%
13	1.51189	0.06852	77.5%
14	1.50674	0.09389	78.0%
15	1.51119	0.09009	77.6%